Geokodierung mit HTTP-Anforderungen

Nikolaos Sitaridis
Universität Ulm
Institut für Epidemiologie
Helmholtzstraße 22
89081 Ulm
nikolaos.sitaridis@uni-ulm.de

Gisela Büchele
Universität Ulm
Institut für Epidemiologie
Helmholtzstraße 22
89081 Ulm
gisela.buechele@uni-ulm.de

Jon Genuneit
Universität Ulm
Institut für Epidemiologie
Helmholtzstraße 22
89081 Ulm
jon.genuneit@uni-ulm.de

Zusammenfassung

Hintergrund:
Unter Geokodierung versteht man die Konvertierung von Adressen in geografische Koordinaten. Geokodierung wird oft benötigt um Daten auf geografischen Karten darzustellen oder auch um geografische Distanzen zu berechnen. Einen freien Service der Geokodierung bietet z.B. Google Maps (www.maps.google.de).

Methodik:
Mittels eines in SAS generierten FILENAME Statement wird auf eine URL zugegriffen, die eine HTTP-Anforderung an Google-Maps stellt. Auf diese HTTP-Anforderung wird eine einzeilige Antwort mit den geografischen Daten im *.csv*-Format generiert, die mit dem INFILE Statement in eine SAS Datendatei eingelesen wird. Mit den erhaltenen geografischen Breiten- und Längengraden wird eine weitere HTTP-Anfrage an EarthTools™ gestellt, um zusätzlich die Höhe über dem Meeresspiegel zu erhalten.

Ergebnisse:
Zur Testung der Anwendung standen insgesamt 1.578 Adressen von Grundschulen in vier deutschsprachigen Regionen (Baden-Württemberg, Bayern, Schweiz, Österreich) und in Polen zur Verfügung. Die Laufzeit für die Geokodierung betrug zwischen 20 und 30 Minuten für den gesamten Beispieldatensatz.

Schlüsselwörter: Geokodierung, Hypertext Transfer Protocol (HTTP), Uniform Resource Locator (URL), Google Maps, Maps-API, EarthTools™, FILENAME, INFILE, Comma Separated Value (CSV)

1 Hintergrund

In vielen Bereichen, welche in der Informationsverarbeitung auf der Tagesordnung stehen, kommt es zur Verwaltung und Verarbeitung von Massendaten. Darunter existieren natürlich auch Adressdaten. In diesem Zusammenhang kommt die Geokodierung, ein Vorgang, welcher Adressdaten in geografische Daten konvertiert und zur Positionierung auf Karten verwendet werden kann, zum Einsatz. Um jedoch Geokodierung durchführen zu können, ist der Zugang zu einer Datenbank mit Geodaten notwendig, der oft mit zusätzlichen Kosten verbunden ist. Google bietet einen Service über Ihre Maps-API an, mit der es möglich ist, Geokodierung für Längen- und Breitengrade via HTTP-Anforderungen durchzuführen. Die Elevation als zusätzliche Information kann über einen Dienst von EarthTools™ erhalten werden.

2 Methodik

In unserem Beispiel stammen die Adressdaten aus einer Microsoft Excel Datei. Aus den Adressdaten muss ein standardisierter Uniform Resource Locator (URL) für die HTTP-Anforderung generiert werden. Dabei ist zu beachten, dass Sonderzeichen vermieden werden. Da mit größter Wahrscheinlichkeit jede einzelne Adresskonstante, wie Straßenname oder Postleitzahl in einer separaten Variable gehalten werden, ist es möglich, mit der SAS Funktion „cat()“ einen String zu generieren der im weiteren Verlauf aufgerufen wird, um die geografischen Daten zu erhalten.

Um auf den Geokodierer des Maps-API zuzugreifen, wird eine Anfrage an http://maps.google.com/maps/geo? mit folgenden Parametern gesendet.

- *q* – jene Adresse die geokodiert werden soll (erforderlich)
- *key* – der API Schlüssel (erforderlich)
- *sensor* – gibt an, ob die Geokodierungsanfrage von einem Gerät mit einem Standortsensor kommt. Dieser Wert muss *true* oder *false* sein (erforderlich)
- *output* – das Format in dem die Ausgabe generiert werden soll. Als Optionen können *xml*, *csv*, *json* (als Standard) und *kml* angewendet werden (erforderlich)
- *gl* – „Optional“ kann der Ländercode angegeben werden

Einige Parameter sind zwingend erforderlich, während andere optional sind.

Syntax zur Stringmanipulation:

```
libname ine excel 'C:\KSFE\ksfe.xls';
data arena; set ine.'Adressen$'n;
     url1=cat("'", 'http://maps.google.com/maps/geo?q=', street,
              '+', number, '+', plz, '+', city,
              '&output=csv&sensor=false&key=&gl=', country,
              "'"));
run;
libname ine clear;
```

Jedem Parameter der URL folgt ein Gleichheitszeichen, um die anschließenden Werte zu übergeben. Ein Beispielaufruf, der die Koordinaten der Gagfah-Arena in Heidenheim an der Brenz anfordert, sieht im Folgenden so aus:
http://maps.google.com/maps/geo?q=schlosshausstrasse+160+89522+heidenheim
&output=csv&sensor=false&key=&gl=de

Die Antwort auf diese HTTP-Anforderung wird in nachstehender Abbildung 1 dargestellt. Hierbei wird das Ergebnis durch vier Kommata getrennte Zahlen dargestellt.

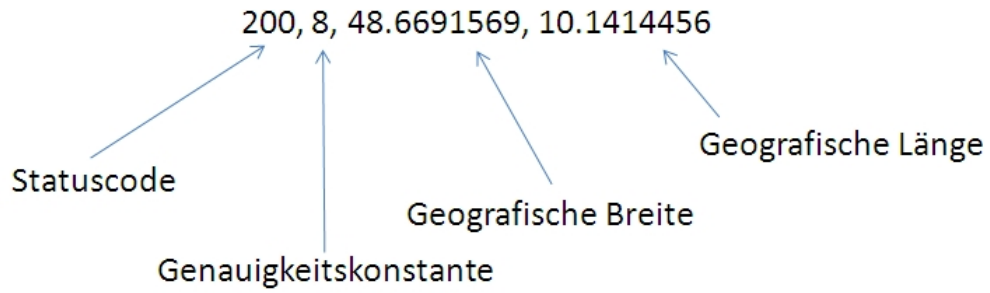

Abbildung 1: Beispielantwort einer HTTP-Anforderung

Um bewusst eine derart kurze Antwort zu erhalten (die im Anschluss übersichtlicher analysiert werden kann) ist die Ausgabe als *csv* zu verwenden.

Der erste Wert liefert den HTTP-Statuscode zurück, siehe hierzu Tabelle 1[1].

Tabelle 1: Statuscodes

Konstante	**Beschreibung**
200	Keine Fehler aufgetreten, Adresse erfolgreich analysiert.
400	Eine Routenanforderung konnte nicht erfolgreich analysiert werden.
500	Eine Geokodierungsanforderung konnte nicht erfolgreich verarbeitet werden, da der genaue Grund für den Fehler nicht bekannt ist.
601	Der HTTP-Parameter q fehlt oder enthält kein Wert. Hat eine leere Adresse zur Folge.
602	Es konnte keine geografische Position für die entsprechende Adresse gefunden werden (relativ neue oder falsche Adresse).
603	Geocode kann aus vertraglichen oder rechtlichen Gründen nicht angegeben werden.
610	Der angegebene Schlüssel ist ungültig oder passt nicht zur Domain.
620	Der angegebene Schlüssel hat innerhalb des Zeitraums von 24 Stunden das Limit für Anforderungen überschritten oder zu viele Anforderungen in einem zu kurzen Zeitraum übermittelt.

Der zweite Wert gibt eine Genauigkeitskonstante der Anfrage an. Für die Genauigkeit gibt es festgeschriebene Konstanten, die in nachstehender Tabelle 2 beschrieben wer-

den. Der dritte Wert liefert die geografische Breite und der letzte Wert gibt die geografische Länge zurück.

Tabelle 2: Genauigkeitskonstanten

Konstante	Beschreibung
0	Unbekannter Ort
1	Land
2	Region (Bundesland, Provinz, Präfektur usw.)
3	Kreis (Bezirk, Gemeinde usw.)
4	Ortschaft (Stadt, Dorf)
5	Postleitzahl (PLZ)
6	Straße
7	Kreuzung
8	Adresse
9	Grundstück (Name des Gebäudes, Einkaufszentrum usw.)

Diese Genauigkeitskonstanten dienen als Maßstab. Sie spiegeln weder ein Ranking noch die Zuverlässigkeit des Ergebnisses wider [2]. Sobald ein Client eine HTTP-Anfrage stellt, versucht der Host die beste Konstante ausfindig zu machen. Je nach der Menge der für den Bereich vorliegenden Daten kann möglicherweise ein bestimmtes Gebäude identifiziert werden, das infrage kommt. Es ist aber auch möglich, dass nur die Straße oder der Bezirk identifiziert werden kann.

Kann nur eine übergeordnete Konstante, wie der Bezirk, zugeordnet werden, wird dennoch eine Breiten-/Längenangabe zurückgegeben. Diese entspricht jedoch dem Zentrum dieses Features und somit wahrscheinlich nicht genau der Position der genauen angegebenen Adresse. Die Genauigkeit informiert Sie über den Maßstab des gefundenen Objekts, sodass Sie wissen, wie akkurat der Geocode wahrscheinlich ist [2].

2.1 Ablaufschema

In Abbildung 2 wird der Ablauf des entwickelten SAS-Makros dargestellt. Der zu öffnende SAS Datensatz ist in Kapitel 2 Methodik erklärt. Dabei muss der SAS Datensatz eine ID zum späteren Zuordnen der Ergebnisse enthalten.
Der Datensatz wird zeilenweise abgearbeitet. In jedem Durchlauf wird die URL über das FILENAME Statement zugewiesen und jeweils mit dem INFILE Statement angesteuert. Zusätzlich wurde ein weiterer Dienst verwendet, das sogenannte EarthTool™. Damit ist es möglich, zu den angeforderten Breiten- und Längengraden die Elevation über den Meeresspiegel zu bekommen. Die Antworten und Fehlermeldungen werden in Makrovariablen gespeichert. Auch Rückmeldungen des Servers werden berücksichtigt. Bei unbekannter Adresse wird eine gekürzte Anfrage (z.B. nur Ort) gestellt. Die Inhalte der Makrovariablen werden mittels PROC SQL in eine Datentabelle abgelegt.

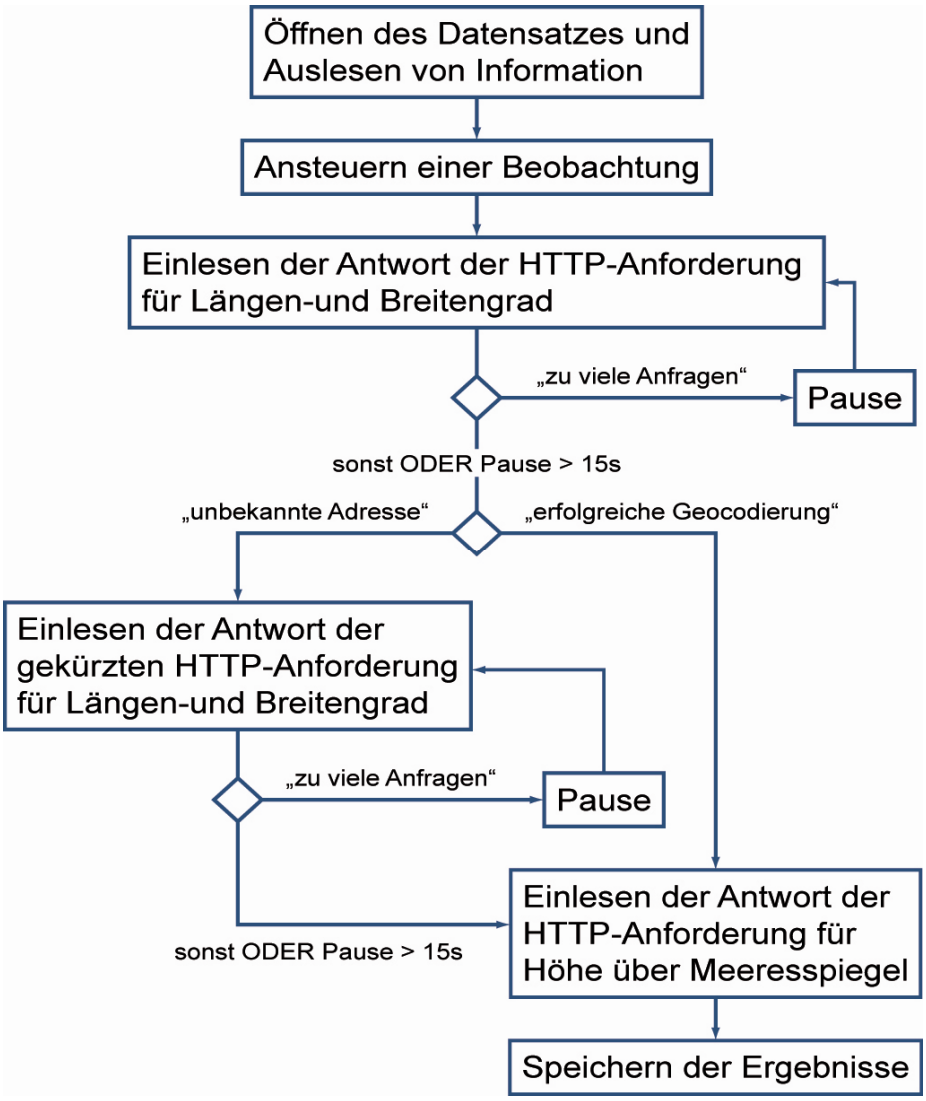

Abbildung 2: Flussdiagramm des SAS-Makros

3 Anwendungsergebnisse

Die Anwendung wurde an einem Beispieldatensatz mit insgesamt 1.578 Adressen von Grundschulen in vier deutschsprachigen Regionen (Baden-Württemberg, Bayern, Schweiz, Österreich) sowie in Polen getestet. Die Laufzeit für die Geokodierung betrug zwischen 20 und 30 Minuten für den gesamten Beispieldatensatz.

Wie in Tabelle 3 zu sehen ist, erzielte die Anwendung auf polnische Adressen schlechtere Ergebnisse als die Anwendung auf deutschsprachige Adressen. Dabei muss berücksichtigt werden, dass die Adressen des Beispieldatensatzes möglicherweise noch Fehler enthalten oder eine sinnvollere Kürzung der Anfrage möglich ist. Laut Google sind bis dato 208 Länder für die Geokodierung verfügbar [3].

Tabelle 3: Geokodierung von N=1.578 Adressen (Angaben in n(%))

Region	erfolgreiche Geokodierung	
	exakt	gekürzte Anforderung
Baden-Württemberg	393 (99.5)	2 (0.5)
Bayern	289 (100)	6 (2.5)
Schweiz	233 (97.5)	6 (2.5)
Österreich	369 (98.7)	5 (1.3)
Polen [1]	223 (79.4)	21 (7.4)

Diskussion

Die Geokodierung unter Verwendung von HTTP-Anforderungen z.B. an Google-Maps und EarthTools™ ist eine kostengünstige Alternative zur Beschaffung von aktuellen, kommerziellen Datenbanken. Dabei müssen die Nutzungsbedingungen der Anbieter selbstverständlich berücksichtigt werden [4]. Die Anwendung an einem Beispieldatensatz zeigte, dass die Lösung in relativ kurzer Zeit eine angemessene Zahl von Adressen bearbeiten kann. Die Qualität der Geokodierung kann nur so gut sein, wie die Qualität der vorliegenden Adressen bzw. die Qualität der Datengrundlage des Anbieters. Für die Zielsetzung unserer Anwendung war die Qualität gut.

Literatur

[1] http://code.google.com/intl/de-DE/apis/maps/documentation/reference.html#GGeoStatusCode (27 April 2010)

[2] http://code.google.com/intl/de-DE/apis/maps/faq.html#geocoder_accuracy (27 April 2010)

[3] http://gmaps-samples.googlecode.com/svn/trunk/mapcoverage_filtered.html (27 April 2010)

[4] http://code.google.com/intl/de-DE/apis/maps/terms.html (27 April 2010)

Anhang A: SAS-Makro zur Geokodierung mittels HTTP-Anforderungen

Das komplette SAS-Makro finden Sie im deutschsprachigen SAS-Wiki unter http://de.saswiki.org/wiki/KSFE2010.

[1] In Polen konnten 37 (13.2%) Adressen nicht geokodiert werden